Competition Highlights

Trevor Gale[1], Erich Elsen[2], Sara Hooker[1], Olivier Temam[2], Scott Gray[3],
Jongsoo Park[4], Cliff Young[1], Utku Evci[1], Niki Parmar[1], Ashish Vaswani[1].
[1]Google, [2]DeepMind, [3]OpenAI, [4]Facebook

micronet-challenge.github.io

Regularization Tricks

The **KAIST AI** team finished 2nd in CIFAR-100 with no quantization. Their model used sparsity along with two tricks: *orthonormal regularization* and *adaptive label smoothing*.

Spectral Restricted Isometry Property Regularization[1]

$$\sigma(W^T W - I) = \sup_{z \in R^n, z \neq 0} \left| \frac{\|Wz\|^2}{\|z\|^2} - 1 \right|$$

Team reported 6% top-1 accuracy gain after applying to all 1x1 convolutions in EfficientNet-like architecture.

Adaptive Label Smoothing

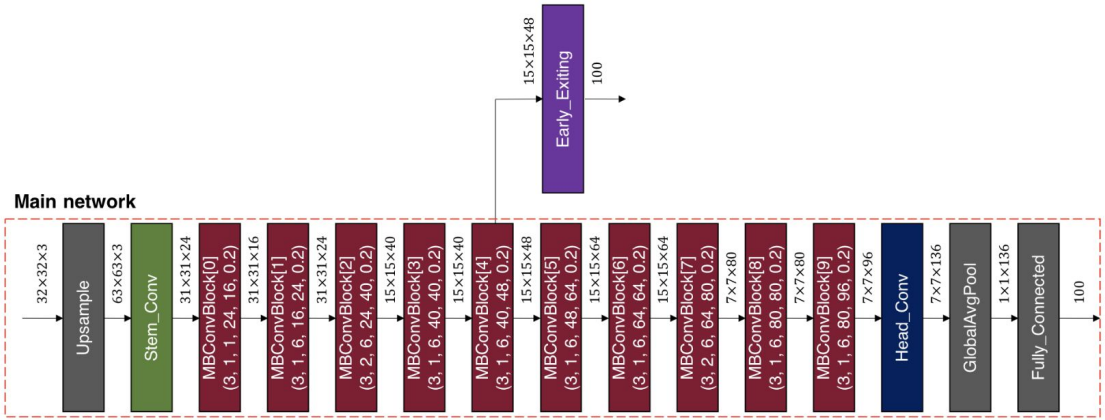
$$c_i = \begin{cases} 1 - \epsilon, & i = k \\ \epsilon \times \dfrac{s(w_i, w_k)}{\sum_{j \neq k} exp|s(w_j, w_k)|}, & i \neq k \end{cases}$$

s(x, y) is cosine similarity, $\mathbf{w}_i$ is weights from final fully-connected layer for class **i**. Adapt label smoothing to take class correlation into account.

[1] arxiv:1810.09102

Early Exiting

For CIFAR-100, the **OSI-AI** team introduced a mechanism for early exiting during inference. Able to terminate inference early on **30%** of examples.



If confidence of early exiting module exceeds threshold the remainder of the main network is not executed. OSI-AI finished 3rd for CIFAR-100.

Ternary Quantization

Used by a number of entries[1]. Can be thought of as a sparse, binary neural network. For both image tasks, team **HHI-MAL** uses decomposition from Trained Ternary Quantization to further avoid multiplications:

Three weight values

$$+W_p,\ 0,\ -W_n$$

Sparse matmul for each nonzero

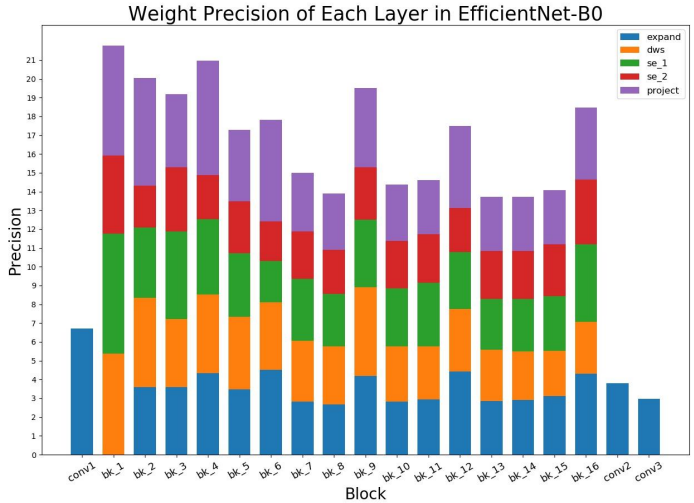
$$W_p * \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}_{W_p} * \begin{bmatrix} 3.2 \\ 0.7 \\ 1.4 \end{bmatrix} + W_n * \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{W_n} * \begin{bmatrix} 0.3 \\ 4.5 \\ 2.8 \end{bmatrix}$$

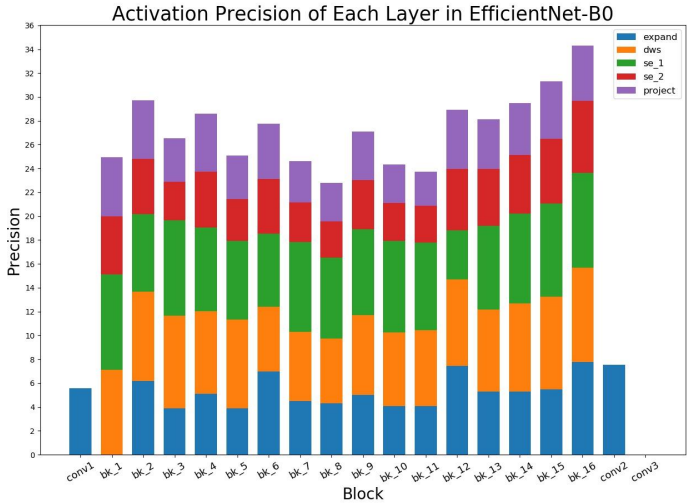
Only **2** multiplications required per output.

[1]Entries MSUNet-V3, HHI-MAL, and ProxylessNAS-TTQ. MB-PM Research uses binary weights.

Adjustable Quantization

For ImageNet, the Texas-EIC team performed quantization aware training with learned channel-wise range and precision factors.





Achieved 75% top-1 with average of **2.94** bits and **4.87** bits for weights and activations respectively.

Complete Leaderboard

micronet-challenge.github.io/leaderboard